# Extensive recombination challenges the utility of *Sugarcane mosaic virus* phylogeny and strain typing

**Luke Braidwood**[1,*], **Sebastian Y. Müller**[1], **and David Baulcombe**[1]

[1]University of Cambridge, Department of Plant Sciences, Cambridge, CB2 3EA, United Kingdom
[*]braidwoodluke@gmail.com

## ABSTRACT

*Sugarcane mosaic virus* (SCMV) is distributed worldwide and infects three major crops: sugarcane, maize, and sorghum. In common with other potyviruses, the impact of SCMV is increased by its interaction with *Maize chlorotic mottle virus* which causes the synergistic maize disease maize lethal necrosis. Here, we characterised maize lethal necrosis-infected maize from multiple sites in East Africa, and found that SCMV was present in all thirty samples. This distribution pattern indicates that SCMV is a major partner virus in the East African maize lethal necrosis outbreak. Consistent with previous studies, our SCMV isolates were highly variable with several statistically supported recombination hot- and cold-spots across the SCMV genome. The recombination events generate conflicting phylogenetic signals from different fragments of the SCMV genome, so it is not appropriate to group SCMV genomes by simple similarity.

## Introduction

*Sugarcane mosaic virus* (SCMV) is a positive-sense single-stranded RNA virus in the *Potyviridae* family (genus *potyvirus*), the largest and most economically damaging family of plant viruses[1]. SCMV can infect three major crops: sorghum, sugarcane (10-35 % yield loss), and maize (20-50 % yield loss), and is thought to be one of the top-ten most economically damaging plant viruses[2–4]. It has been reported in 84 countries across the 6 inhabited continents and this cosmopolitan distribution is likely due to worldwide trade in its host crops for hundreds of years (fig. 1)[5]. Most potyviruses, including SCMV, are spread non-persistently by aphid species[6] but SCMV can also spread via movement of infected root cane (sugarcane) and through maize seeds and pollen[7,8].

The *potyvirus* genus is notable for its size (>150 species) and extensive involvement in synergistic plant viral conditions. Typically, potyviruses enhance the titre of the partner virus in synergistic interactions through a process that is dependent on the multifunctional helper-component protease (HC-Pro)[9,10]. Synergism between potyviruses, including SCMV, and *Maize chlorotic mottle virus* (MCMV) causes maize lethal necrosis (MLN) that can cause total yield loss[11]. SCMV threatens both food security and economic development because maize and sorghum are vital staple foods, while sugarcane is an important cash crop. Despite SCMV being present in East Africa and China for decades, its impact in both regions has been enhanced by the recent arrival of MCMV, and therefore MLN[12–16]. Increased understanding of the variability and evolution of SCMV in these regions may inform future disease control measures.

SCMV has a typical *potyvirus* genome: a roughly 9.5 kb monopartite positive-sense single-stranded RNA molecule (fig. 1b) which is packaged into around 2,000 helically arranged coat protein (CP) monomers to form flexuous virions 750 nm long and 13 nm wide. The 5' end of the genome is capped by the 25 kDa Vpg protein, and the 3' end is poly-adenylated. Translation of the genome produces a single polyprotein which is cleaved by three viral-encoded proteases to generate ten multifunctional proteins[1,17]. An additional protein, P3N-PIPO, is generated due to transcriptional slippage in the P3 gene at a conserved GAAAAAA motif during genome replication[18,19].

*Potyviridae* evolution features extensive intra-specific recombination[20,21], which likely occurs when the viral RNA-dependent RNA polymerase (RdRP) switches between viral genome templates[22] during virus replication. Reported recombination hot-spots are in the P1 region of *Turnip mosaic virus* and in the CI-NIa-protease region of several species (fig. 1)[23–34]. Predicted recombination breakpoints in the SCMV genome are in CI, NIb, NIa-VPg, and NIa-Pro, and the 6K1-VPg-NIa-Pro-NIb region has been called a recombination hot-spot, although without statistical support[28–34]. Recombination complicates phylogenetic analyses because various genome regions in a single individual may have different evolutionary histories. Accordingly, constructing phylogenies using different sections of SCMV and other potyviral genomes produces conflicting trees[35,36]. Recombination may also impede virus detection because increased genomic variation may lead to false negative results with common techniques such as PCR and antibody ELISA[11,37].

There are multiple potyviruses present in East Africa which could act as partner viruses to MCMV. Therefore, we decided

to survey MLN-infected maize in Kenya and Ethiopia using next-generation sequencing (NGS) to allow identification and analysis of the partner viruses in this region. The only partner virus we detected was SCMV, and these data were then used to look for signals of historical recombination in the SCMV genome. We also assessed the suitability of traditional phylogenetic methods for SCMV genomic data.

## Results and Discussion

### Sequencing of MLN-infected maize reveals SCMV

In August 2014 we collected 23 MLN-symptomatic (mosaic and chlorosis on leaves) maize samples from 13 Kenyan and 4 Ethiopian sites (table S1) and performed NGS RNA-seq (ArrayExpress accession: E-MTAB-7002). All samples contained both MCMV, characterised previously[38], and SCMV (fig. S1). The 23 assembled SCMV sequences ranged from 2,191 bp to 9,632 bp, which is 23 % to 100 % of the longest previously reported SCMV sequence (available in GenBank: MH093717-MH093739). We aligned and manually trimmed the long (>8000 bp) SCMV contigs to the full SCMV genomes available in GenBank for further analysis.

There were small insertions in the 5' and 3' untranslated regions (UTRs) in one (JX047421.1) and three (JX185303.1, EU091075.1, GU474635.1) isolates respectively but most insertion/deletion variation was over a 200 bp region of the CP coding sequence (fig. 2a, S2, S3). CP indels were not distributed according to the geographic location of isolates. A 15 bp insertion, for example, was present in isolates from Ethiopia (1), Kenya (9), Rwanda (2), USA (1), and Mexico (1) whereas a 3 bp insertion at the same locus is present in Kenya (2), China (10), and Ecuador (1).

We assessed nucleotide polymorphism diversity of multiple sequence alignments using DNAsp5[39]. Diversity was high with 4,289 mutations spread over 2,831 sites and an average of 1121.1 nucleotide differences between sequences. Nucleotide diversity across the genome was 0.17, higher than for most RNA viruses but within the range previously reported for SCMV[30,33]. There were high polymorphism regions in the N-termini of P1 and CP and the most conserved regions were in the central domain of P3 and the 3' UTR (fig. 2b-c). P1 is a serine protease with a known hyper-variable region[40]. The variable region in the CP N-terminus (fig. 2b-c, S2, S3) corresponds to a domain with variable amino acid length and low conservation[41,42], with episodic positive selection detected by MEME analysis (fig. S4)[43,44]. This N-terminal domain is surface located, raising the possibility that variation in this region may alter interactions with host or vector proteins[45,46].

The conserved P3 protein is essential for potyviral replication but it is also the locus of the cryptic fusion protein P3N-PIPO and this overlapping open reading frame is an extra constraint to evolution of the nucleotide sequence[47]. Interestingly, in the P3N-PIPO region, there were also sites with episodic positive selection detected by MEME (fig. S4). The 3' UTR of potyviruses contains a poly-A tail to promote genome stability and translation which is completely conserved.

### Evidence for SCMV recombination

The alignment patterns of several samples suggested recombination, with different regions of the same sample genome showing closest alignment to divergent reference genomes (fig. 3a). To simplify the analysis whilst retaining maximum diversity, we subsampled the alignment of 116 sequences. We generated a nucleotide identity matrix (table S2) and grouped sequences with >99 % similarity, then kept the longest sequence in each group. This produced a final dataset of 55 SCMV genomes/contigs, including 13 from our NGS libraries.

Splits network analysis can detect and visualize conflicting signals from a phylogenetic dataset[48]. Conflicting signals imply that the relationship between sequences is different depending on the part of the sequence being analysed, and they can be caused by recombination or horizontal gene transfer. Splits networks with reticulate shapes rather than bifurcating tree shapes indicate conflicting phylogenetic signals. Here, we found the splits network derived from our SCMV sequences to be very different from a bifurcating tree indicating conflicting phylogenetic signals and implying recombination (fig. 3b). Additional independent evidence for SCMV recombination comes from the distribution of multiple indels that do not correlate with geographic or phylogenetic proximity (fig. 2a, S2, S3).

To estimate the number and location of recombination breakpoints we used Recombination Detection Programme 4 (RDP4) to predict the locations of recombination events in SCMV[49]. RDP uses multiple algorithms to locate sites in an alignment at which phylogenetic signals change rapidly, which is indicative of a recombination event. The recombination scheme suggested multiple recombination events, with many between geographic regions (fig. 3c-d). There was notable reciprocal exchange of recombinant fragments between European and Chinese isolates (fig. 3c-d), and between Chinese and African isolates. There was also evidence for intra-region recombination in the regions with more than five isolates (China and Africa). Recombination was more frequent between strains within a region than between strains in different region, as expected. Additionally, there were 28 recombination events with unknown parents (i.e. genome fragments not closely related to any known isolates), demonstrating that more sequencing data will be required to fully describe worldwide recombination patterns.

To search for regions of the genome with an over- or under-representation of recombination, we counted the recombination breakpoints in sliding windows across the SCMV genome (fig. 4a), calculated likelihood ratios, and used permutation testing to

identify statistically significant regions (fig. 4b). The permutation test randomly places recombinant fragments spanning the same number of variable nucleotide positions as each detected recombinant fragment, which controls for sequence variability and generates a density map of where recombination is more likely to be detected in the SCMV alignment.

'Global hot/cold-spots' were defined as those with more breakpoints than 95 % of the sliding windows across the genome, and 'local hot/cold-spots' were those with more breakpoints than 99 % of sliding windows at that position. Global recombination hot-spots were present at the 5' and 3' genomic termini. We detected nine local recombination hot-spots, and twelve local recombination cold-spots (fig. 4). The hot-spots were concentrated in the 3' region encoding NIb and CP, with single hot-spots in *CI*, *P3N-PIPO*, and the *P1/HC-Pro* junction (fig. 4c). The cold-spots were distributed more uniformly across the first 7500 bp of the genome. These are the first statistically supported recombination hot- and cold-spots reported in SCMV.

Recombination can promote nucleotide diversity by mixing lineages, and we note that the 3' genomic region encoding CP in SCMV has high nucleotide diversity (fig. 2), and a concentration of recombination hot-spots (fig. 4). *Potyvirus* recombination hot-spots have previously been observed in the C-terminal region of *CI*, which we also observed, and in P1, which we did not[23–27]. Recombination is clearly a major force in SCMV evolution (fig. 3), as in the *Potyviridae* generally[20,50].

## Is making an SCMV phylogeny useful?

The purpose of a phylogeny is to describe the evolutionary history of biological entities. This exercise has academic value, in tracing the history of life, and practical value, in organising similar biological entities into clades. If a phylogeny does not describe evolutionary history, and does not group biological entities into self-similar clades, it is an inappropriate analysis (due to the methodology or the underlying data) and does not contain useful information. Given the high levels of recombination between our SCMV sequences, we decided to investigate whether further phylogenetic analysis is appropriate.

There are many published SCMV phylogenies, based on CP sequences and whole genomes, which place isolates into two to six strains with variable names, see Gao *et al.* (2011) for a helpful summary[51]. Whole genome phylogenies of SCMV group isolates into four strains (IA, IB, II, III, IV), with around 80 % nucleotide similarity between strains[33,51]. African SCMV genomes sequenced in this study form two novel clusters (AI and AII) of sequence identity, decreasing the separation between previously reported strains (fig. 3b).

Simulations show that phylogenetic analyses are most severely impacted by recombination when breakpoints occur near the centre of alignments, and by recent recombination between diverged taxa[52]. Our recombination analysis shows evidence of recombination between divergent (<80 % nucleotide identity) SCMV isolates, in the centre of both genomes and CP sequences (fig. 3, S5, table S3). Therefore, there is no single evolutionary history for phylogenetic analyses to infer. To statistically test for conflicting phylogenetic signals, we constructed phylogenies by maximum likelihood using the whole SCMV genome, and three sections of the alignment (section 1: positions 963-2,764 in the original alignment, section 2: 2,875-5,103, and section 3: 5,181-8,036) without recombination hot-spots (fig. 4, supplemental data d1). We chose alignment sections with a minimum number of recombination events (i.e. containing cold-spots) which were separated with recombination hot-spots (fig. 4c). Tree incongruence was tested statistically using a Shimodaira-Hasegawa test (SH-test) for each pair of trees. The log likelihood differences were 12,222 between sections 1 and 2, 12,739 between 1 and 3, and 15,692 between 2 and 3 ($p<1e\text{-}7$ and $n = 2$ trees for all comparisons), confirming significant differences between the trees, which can be visualised using tanglegrams or identity matrices (fig. S6).

## Conclusion

Viral studies often present a phylogeny followed by evidence of extensive recombination, showing that the central assumption of the phylogenetic analysis was violated[23,30,31,33]. Multiple evolutionary histories within a genome are valid and averaging these different histories does not produce the true evolutionary history of the genome[52]. Imposing a bifurcating tree structure on a dataset which does not have a single, bifurcating evolutionary history will introduce systematic error. We argue that in organisms with unknown or high recombination rates, such as RNA viruses, recombination analyses should be performed initially, then used to inform the phylogenetic approach taken, as in Ohshima *et al.* (2007)[25]. Splits network analysis is appropriate for all alignments, but standard phylogenetic methods may not be, depending on the splits network results. Phylogenetic analyses of whole genomes may be desirable to identify viral strains containing isolates with a broadly similar evolutionary history. However, the presence of sequences from different strains which have entered due to recombination may confound phenotyping and molecular attempts at strain identification in the field.

We have shown that SCMV is in complex with MCMV in MLN-infected maize in East Africa, and that producing SCMV phylogenies does not produce useful classification systems or describe biological truth (fig. S6)[52]. We conclude that constructing phylogenetic trees is inappropriate for SCMV due to extensive historical recombination between divergent isolates. This may also have implications for studies of other RNA viruses, and phylogenies of other organisms with high recombination rates. There are multiple avenues for progress in this field - for example the statistical framework for assessing splits networks is not well developed, there are no automated approaches for locating viral recombination hot-spots, and visualisation of reticulate recombination networks .

## Methods

### NGS of MLN-infected maize

We collected maize leaf samples from Kenya and Ethiopia in August 2014 (table S1), storing samples in RNA-later (Ambion) on dry ice. To extract RNA, we used Trizol (Ambion) according to manufacturer's instructions. We depleted ribosomal RNA with the Ribo-Zero Magnetic Kit (Plant Leaf - Epicentre). To generate indexed stranded libraries, we used Scriptseq V2 RNA-Seq Library Preparation kits and Scriptseq Index PCR primers (Epicentre). Library concentration and quality were confirmed using Qubit (Life Technologies) and a Bioanalyzer High Sensitivity DNA Chip (Agilent Technologies). Beijing Genomics Institute performed 100 bp paired-end sequencing on one lane of a HiSeq 2000 (Illumina).

### NGS quality control

We used a custom python script to demultiplex the libraries allowing one error in index sequences, then trimmed adaptors using Trim galore! (parameters: –phred64 –fastQC –illumina –length 30 –paired –retain_unpaired input_1.fq input_2.fq)[53,54]. String matching deduplication (deletion of identical reads) was performed using Quality Assessment of Short Read (QUASR) pipeline scripts[55].

### SCMV consensus sequence generation

To generate SCMV genome sequences, we aligned libraries to a bowtie2 reference containing all SCMV genomes available in NCBI GenBank in March 2016 (parameters: *-D 20 -R 2 -N 1 -L 20 -i S,0,2.50 –phred64 –maxins 1000 –fr*)[56]. Next, we extracted SCMV-aligning reads and performed *de novo* assembly using Trinity (v2.0.2), extracted contigs above 2 kb in length, then inspected and curated (if necessary) SCMV contigs[57]. To generate SCMV consensus sequences, we aligned each library to its respective Trinity contig using bowtie2, generated pileups using samtools, and called sequences using the QUASR script *pileup_consensus.py*, with a threshold of zero or ten % of reads for the calling of ambiguity codes (parameters: *–ambiguity 0—10 –dependent –cutoff 25 –lowcoverage 20*)[58].

### SCMV alignment and diversity analysis

SCMV genomes generated in this study were combined with those in GenBank and aligned using MUSCLE (gap extension cost: 800, other settings default) in MEGA6, with separate alignments for genomes called with and without ambiguous bases[59]. We checked the alignments manually in JALview and refined where necessary[60]. To construct a nucleotide identity matrix we used the dist.alignment function from the R package seqinr. We obtained diversity metrics using the alignment without ambiguous base calls in DnaSP v5[39].

### SCMV recombination analysis

Recombination analysis was performed with the alignment containing no ambiguous base calls. To generate splits networks, we used SplitsTree4 using default settings - distances calculated by uncorrected P, and network generated by neighbour-net[48]. To generate more specific predictions of recombination, we used RDP4, using the algorithms RDP, GENECONV, MaxChi, BootScan, and SiScan (all default settings), and reviewed all breakpoints manually. Recombination network diagrams were generated by constructing interaction matrices for regions and SCMV isolates. The interaction matrix was converted into a regional recombination network (fig. 3c) using the ggraph function from the ggraph R package, while the individual interaction networks (fig. 3d) were constructed using the ggnet2 function of the R package GGally.

### Dendrogram and phylogeny construction

The nucleotide identity dendrogram was constructed from the identity matrix using the heatmap.2 function of the gplots R package. Phylogenies were constructed from the alignment without ambiguous bases using RAxML-HPC2 (8.2.10) on XSEDE, hosted by the CIPRES science gateway (parameters: -T 4 -N autoMRE -n result -s infile.txt -m GTRCAT -q part.txt -c 25 -p 12345 -f a -x 12345 –asc-corr lewis)[61]. Phylogenies were compared using the tanglegram function of Dendroscope (3.5.9)[62].

### Statistical analysis

We used an SH-test to test for tree incongruence[63] between phylogenies constructed using the three alignment sections. Trees for each section were generated using the methods above and compared using the SH-test as implemented in the R-package phangorn[64] in a pairwise fashion. To determine whether the SH-test is appropriate for these data we created a negative control dataset from our alignment in which we did not expect tree disagreement. In the negative control dataset, the null hypothesis (of identical tree architectures) should not be rejected. To generate the negative control dataset, we divided a region containing a recombination cold spot (positions 2,224 to 2,586 of the original alignment), which should have had little recombination and therefore have a consistent evolutionary history into two sections (positions 2,224 to 2,400 and 2,401 to 2,586). The log likelihoods of the tree topologies constructed from these sections were -5,895 and -5,735 respectively with a difference of 59.8. Subsequent testing with the SH-test did not reject the null hypothesis of the topologies agreeing (p=0.21).

## Acknowledgements

The authors would like to acknowledge colleagues at KALRO for their assistance in maize sampling, to farmers in all study areas for providing maize samples, and to John Welch for guidance on phylogenetic analyses. L.B. is supported by the BBSRC DTP and the 2Blades Foundation, S.Y.M. was supported by the European Research Council Advanced Investigator Grant ERC-2013-AdG 340642 - TRIBE and D.C.B. is supported by the Royal Society Edward Penley Abraham Research Professorship.

## Author contributions statement

L.B. performed sampling, sequencing, data analysis, and drafted the paper. S.Y.M. performed data analysis, drafted text, and edited the manuscript. D.C.B. designed experiments and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Data availability

RNA-seq data have been deposited in the ArrayExpress database[65] at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7002. SCMV contigs are available in Genbank under accession numbers MH093717-MH093739.

## References

1. López-Moya, J. J., Valli, A. & García, J. A. Potyviridae. In *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, Chichester, 2009).
2. Viswanathan, R. & Balamuralikrishnan, M. Impact of mosaic infection on growth and yield of sugarcane. *Sugar Tech* **7**, 61–65 (2005).
3. Zhu, M. *et al.* Maize Elongin C interacts with the viral genome-linked protein, VPg, of Sugarcane mosaic virus and facilitates virus infection. *The New phytologist* **203**, 1291–304 (2014). DOI 10.1111/nph.12890.
4. Rybicki, E. P. A Top Ten list for economically important plant viruses. *Arch. Virol.* **160**, 17–20 (2015). DOI 10.1007/s00705-014-2295-9.
5. Wu, L., Zu, X., Wang, S. & Chen, Y. Sugarcane mosaic virus – Long history but still a threat to industry. *Crop. Prot.* **42**, 74–78 (2012). DOI 10.1016/j.cropro.2012.07.005.
6. Teakle, D. S. & Grylls, N. E. Four strains of sugarcane mosaic virus infecting cereals and other grasses in Australia. *Aust. J. Agric. Res.* **24**, 465–477 (1973).
7. Li, L., Wang, X. & Zhou, G. Analyses of maize embryo invasion by Sugarcane mosaic virus. *Plant Sci.* **172**, 131–138 (2007). DOI 10.1016/j.plantsci.2006.08.006.
8. Perera, M. F., Filipone, M., Noguera, A. S., Cuenya, M. I. & Castagnaro, A. P. An overview of the sugarcane mosaic disease in south america. *Funct. Plant Sci. Biotechnol.* **6**, 98–107 (2012).
9. Shi, X. M., Miller, H., Verchot, J., Carrington, J. C. & Vance, V. B. Mutations in the region encoding the central domain of helper component-proteinase (HC-Pro) eliminate potato virus X/potyviral synergism. *Virol.* **231**, 35–42 (1997). DOI 10.1006/viro.1997.8488.
10. González-Jara, P. *et al.* A single amino acid mutation in the Plum pox virus helper component-proteinase gene abolishes both synergistic and RNA silencing suppression activities. *Phytopathol.* **95**, 894–901 (2005). DOI 10.1094/PHYTO-95-0894.
11. Mahuku, G. *et al.* Maize lethal necrosis (MLN), an emerging threat to maize-based food security in sub-Saharan Africa. *Phytopathol.* (2015). DOI 10.1094/PHYTO-12-14-0367-FI.
12. Louie, R. Sugarcane Mosaic Virus in Kenya. *Plant Dis.* **64**, 944 (1980). DOI 10.1094/PD-64-944.
13. Chen, J., Chen, J. & Adams, M. J. Characterisation of potyviruses from sugarcane and maize in China. *Arch. Virol.* **147**, 1237–1246 (2002). DOI 10.1007/s00705-001-0799-6.
14. Wangai, A. W. *et al.* First report of Maize chlorotic mottle virus and maize lethal necrosis in Kenya. *Plant Dis.* **96**, 1582–1582 (2012). DOI 10.1094/PDIS-06-12-0576-PDN.

**15.** Xie, L. *et al.* Characterization of Maize chlorotic mottle virus associated with maize lethal necrosis disease in China. *J. Phytopathol.* **159**, 191–193 (2011). DOI 10.1111/j.1439-0434.2010.01745.x.

**16.** Achon, M., Serrano, L., Clemente-Orta, G. & Sossai, S. First report of Maize chlorotic mottle virus on a perennial host, Sorghum halepense, and maize in Spain. *Plant Dis.* **101**, 393 (2017).

**17.** Urcuqui-Inchima, S., Haenni, A.-L. & Bernardi, F. Potyvirus proteins: a wealth of functions. *Virus Res.* **74**, 157–175 (2001). DOI 10.1016/S0168-1702(01)00220-9.

**18.** Olspert, A., Chung, B. Y.-W., Atkins, J. F., Carr, J. P. & Firth, A. E. Transcriptional slippage in the positive-sense RNA virus family Potyviridae. *EMBO reports* e201540509 (2015). DOI 10.15252/embr.201540509.

**19.** Rodamilans, B. *et al.* RNA polymerase slippage as a mechanism for the production of frameshift gene products in plant viruses of the Potyviridae family. *J. virology* **89**, 6965–6967 (2015). DOI 10.1128/JVI.00337-15.

**20.** Chare, E. R. & Holmes, E. C. A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch. Virol.* **151**, 933–946 (2006). DOI 10.1007/s00705-005-0675-x.

**21.** Sztuba-Solińska, J., Urbanowicz, A., Figlerowicz, M. & Bujarski, J. J. RNA-RNA recombination in plant virus replication and evolution. *Annu. review phytopathology* **49**, 415–43 (2011). DOI 10.1146/annurev-phyto-072910-095351.

**22.** Bujarski, J. J. Genetic recombination in plant-infecting messenger-sense RNA viruses: overview and research perspectives. *Front. plant science* **4**, 68 (2013). DOI 10.3389/fpls.2013.00068.

**23.** Seo, J.-K. *et al.* Molecular variability and genetic structure of the population of soybean mosaic virus based on the analysis of complete genome sequences. *Virol.* **393**, 91–103 (2009). DOI 10.1016/j.virol.2009.07.007.

**24.** Tugume, A. K., Cuéllar, W. J., Mukasa, S. B. & Valkonen, J. P. T. Molecular genetic analysis of virus isolates from wild and cultivated plants demonstrates that East Africa is a hotspot for the evolution and diversification of Sweet potato feathery mottle virus. *Mol. Ecol.* **19**, 3139–3156 (2010). DOI 10.1111/j.1365-294X.2010.04682.x.

**25.** Ohshima, K. *et al.* Patterns of recombination in turnip mosaic virus genomic sequences indicate hotspots of recombination. *J. Gen. Virol.* **88**, 298–315 (2007). DOI 10.1099/vir.0.82335-0.

**26.** Bousalem, M. *et al.* High genetic diversity, distant phylogenetic relationships and intraspecies recombination events among natural populations of Yam mosaic virus: a contribution to understanding potyvirus evolution. *J. Gen. Virol.* 243–255 (2000).

**27.** Moreno, I. M. *et al.* Variability and genetic structure of the population of watermelon mosaic virus infecting melon in Spain. *Virol.* **318**, 451–460 (2004). DOI 10.1016/j.virol.2003.10.002.

**28.** Achon, M. A., Serrano, L., Alonso-Dueñas, N. & Porta, C. Complete genome sequences of Maize dwarf mosaic and Sugarcane mosaic virus isolates coinfecting maize in Spain. *Arch. Virol.* **152**, 2073–2078 (2007). DOI 10.1007/s00705-007-1042-x.

**29.** Gell, G., Sebestyén, E. & Balázs, E. Recombination analysis of Maize dwarf mosaic virus (MDMV) in the Sugarcane mosaic virus (SCMV) subgroup of potyviruses. *Virus Genes* **50**, 79–86 (2015). DOI 10.1007/s11262-014-1142-0.

**30.** Li, Y., Liu, R., Zhou, T. & Fan, Z. Genetic diversity and population structure of Sugarcane mosaic virus. *Virus Res.* **171**, 242–246 (2013). DOI 10.1016/j.virusres.2012.10.024.

**31.** Moradi, Z., Mehrvar, M., Nazifi, E. & Zakiaghl, M. The complete genome sequences of two naturally occurring recombinant isolates of Sugarcane mosaic virus from Iran. *Virus Genes* **52**, 270–280 (2016). DOI 10.1007/s11262-016-1302-5.

**32.** Padhi, A. & Ramu, K. Genomic evidence of intraspecific recombination in sugarcane mosaic virus. *Virus genes* **42**, 282–5 (2011). DOI 10.1007/s11262-010-0564-6.

**33.** Xie, X. *et al.* Molecular variability and distribution of Sugarcane mosaic virus in Shanxi, China. *PLoS ONE* **11**, 1–12 (2016). DOI 10.1371/journal.pone.0151549.

**34.** Zhong, Y. *et al.* Identification of a naturally occurring recombinant isolate of Sugarcane mosaic virus causing maize dwarf mosaic disease. *Virus Genes* **30**, 75–83 (2005). DOI 10.1007/s11262-004-4584-y.

**35.** Handley, J. A., Smith, G. R., Dale, J. L. & Harding, R. M. Sequence diversity in the CP coding region of eight sugarcane mosaic potyvirus isolates infecting sugarcane in Australia. *Arch. virology* **143**, 1145–1153 (1998). DOI 10.1007/BF01718631.

**36.** Mishra, R., Patil, S., Patil, A. & Patil, B. L. Sequence diversity studies of papaya ringspot virus isolates in south india reveal higher variability and recombination in the 5'-terminal gene sequences. *VirusDisease* (2019). DOI 10.1007/s13337-019-00512-x.

**37.** Adams, I. P. *et al.* Use of next-generation sequencing for the identification and characterization of Maize chlorotic mottle virus and Sugarcane mosaic virus causing maize lethal necrosis in Kenya. *Plant Pathol.* **62**, 741–749 (2013). DOI 10.1111/j.1365-3059.2012.02690.x.

**38.** Braidwood, L. *et al.* Maize chlorotic mottle virus exhibits low divergence between differentiated regional sub-populations. *Sci. Reports* **8**, 1–9 (2018). DOI 10.1038/s41598-018-19607-4.

**39.** Librado, P. & Rozas, J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinforma.* **25**, 1451–1452 (2009). DOI 10.1093/bioinformatics/btp187.

**40.** Pasin, F., Simón-Mateo, C. & García, J. A. The hypervariable amino-terminus of P1 protease modulates potyviral replication and host defense responses. *PLoS Pathog.* **10** (2014). DOI 10.1371/journal.ppat.1003985.

**41.** Frenkel, M. J. *et al.* Unexpected sequence diversity in the amino-terminal ends of the coat proteins of strains of sugarcane mosaic virus. *J. Gen. Virol.* **72**, 237–242 (1991). DOI 10.1099/0022-1317-72-2-237.

**42.** Xiao, X. W., Frenkel, M. J., Teakle, D. S., Ward, C. W. & Shukla, D. D. Sequence diversity in the surface-exposed amino-terminal region of the coat proteins of seven strains of sugarcane mosaic virus correlates with their host range. *Arch. Virol.* 399–408 (1993).

**43.** Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS genetics* **8**, e1002764 (2012).

**44.** Kosakovsky Pond, S. L. & Frost, S. D. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. biology evolution* **22**, 1208–1222 (2005).

**45.** Shukla, D. D., Strike, P. M., Tracy, S. L., Gough, K. H. & Ward, C. W. The N and C termini of the coat proteins of potyviruses are surface-located and the N terminus contains the major virus-specific epitopes. *J. Gen. Virol.* **69**, 1497–1508 (1988). DOI 10.1099/0022-1317-69-7-1497.

**46.** López-Moya, J. J., Wang, R. Y. & Pirone, T. P. Context of the coat protein DAG motif affects potyvirus transmissibility by aphids. *J. Gen. Virol.* **80**, 3281–3288 (1999). DOI 10.1099/0022-1317-80-12-3281.

**47.** Firth, A. E. & Brown, C. M. Detecting overlapping coding sequences in virus genomes. *BMC Bioinforma.* **6**, 1–6 (2006). DOI 10.1186/1471-2105-7-75.

**48.** Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006). DOI 10.1093/molbev/msj030.

**49.** Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, 1–5 (2015). DOI 10.1093/ve/vev003.

**50.** Revers, F., Le Gall, O., Candresse, T., Le Romancer, M. & Dunez, J. Frequent occurrence of recombinant potyvirus isolates. *J. Gen. Virol.* **77**, 1953–1965 (1996). DOI 10.1099/0022-1317-77-8-1953.

**51.** Gao, B., Cui, X.-W., Li, X.-D., Zhang, C.-Q. & Miao, H.-Q. Complete genomic sequence analysis of a highly virulent isolate revealed a novel strain of Sugarcane mosaic virus. *Virus genes* **43**, 390–7 (2011). DOI 10.1007/s11262-011-0644-2.

**52.** Posada, D. & Crandall, K. The effect of recombination on the accuracy of phylogeny estimation. *J. molecular evolution* **54**, 396–402 (2002). DOI 10.1007/s00239.

**53.** Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads (2011).

**54.** Krueger, F. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files (2015).

**55.** Watson, S. J. *et al.* Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philos. transactions Royal Soc. Lond. Ser. B, Biol. sciences* **368**, 20120205 (2013). DOI 10.1098/rstb.2012.0205.

**56.** Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012). DOI 10.1038/nmeth.1923.

**57.** Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. biotechnology* **29**, 644–52 (2011). DOI 10.1038/nbt.1883.

**58.** Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma. (Oxford, England)* **27**, 2987–93 (2011). DOI 10.1093/bioinformatics/btr509.

**59.** Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. biology evolution* **30**, 2725–9 (2013). DOI 10.1093/molbev/mst197.

**60.** Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinforma.* **20**, 426–427 (2004). DOI 10.1093/bioinformatics/btg430.

**61.** Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gatew. Comput. Environ. Work. GCE 2010* (2010). DOI 10.1109/GCE.2010.5676129.

**62.** Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012). DOI 10.1093/sysbio/sys062.

**63.** Planet, P. J. Tree disagreement: measuring and testing incongruence in phylogenies. *J. biomedical informatics* **39**, 86–102 (2006).

**64.** Schliep, K. P. phangorn: phylogenetic analysis in r. *Bioinforma.* **27**, 592–593 (2010).

**65.** Kolesnikov, N. *et al.* Arrayexpress update—simplifying data submissions. *Nucleic acids research* **43**, D1113–D1116 (2014).

**66.** Wickham, H., Chang, W. *et al.* ggplot2: An implementation of the grammar of graphics. *R package version 0.7, URL: http://CRAN. R-project. org/package= ggplot2* (2008).

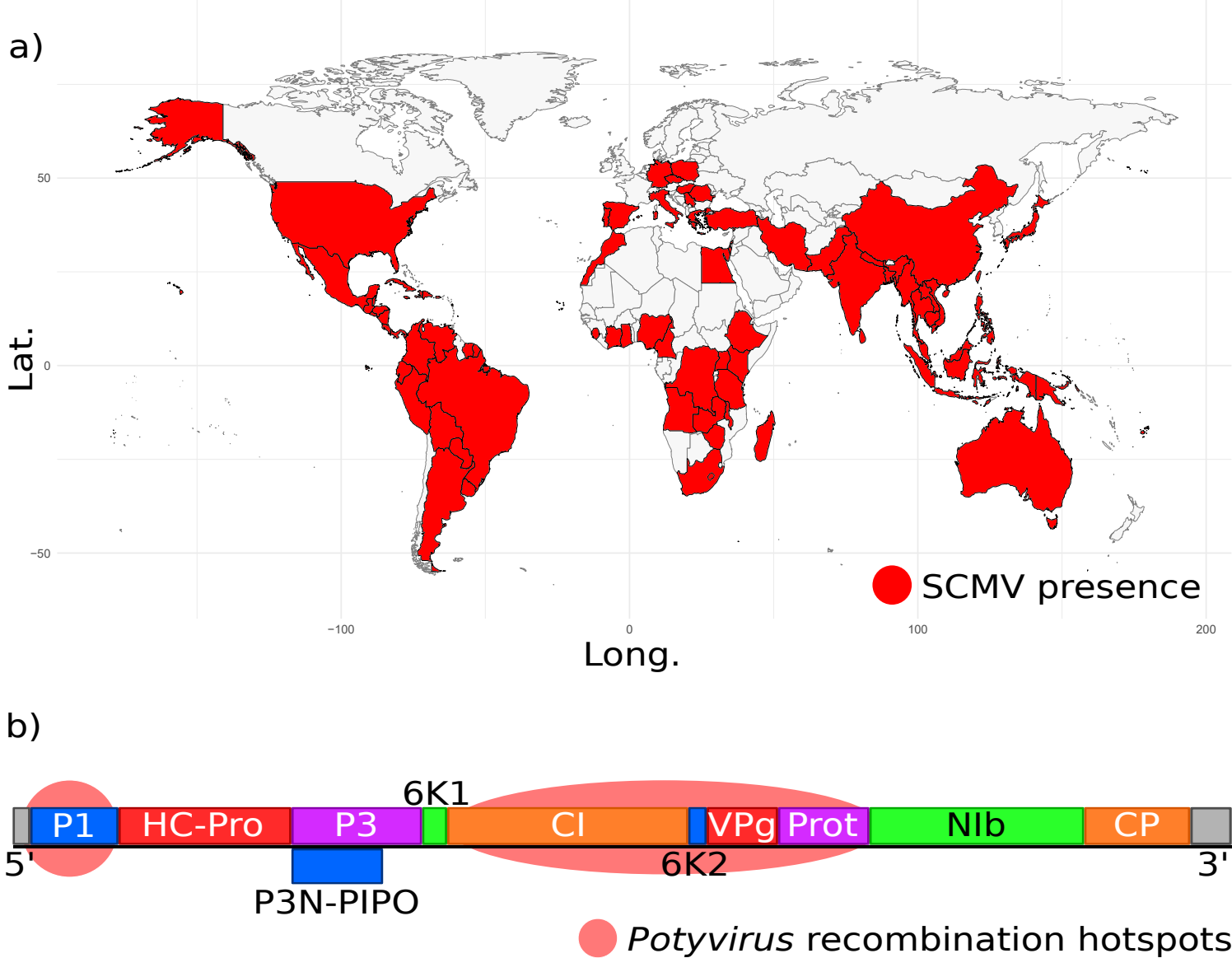
## Figure captions

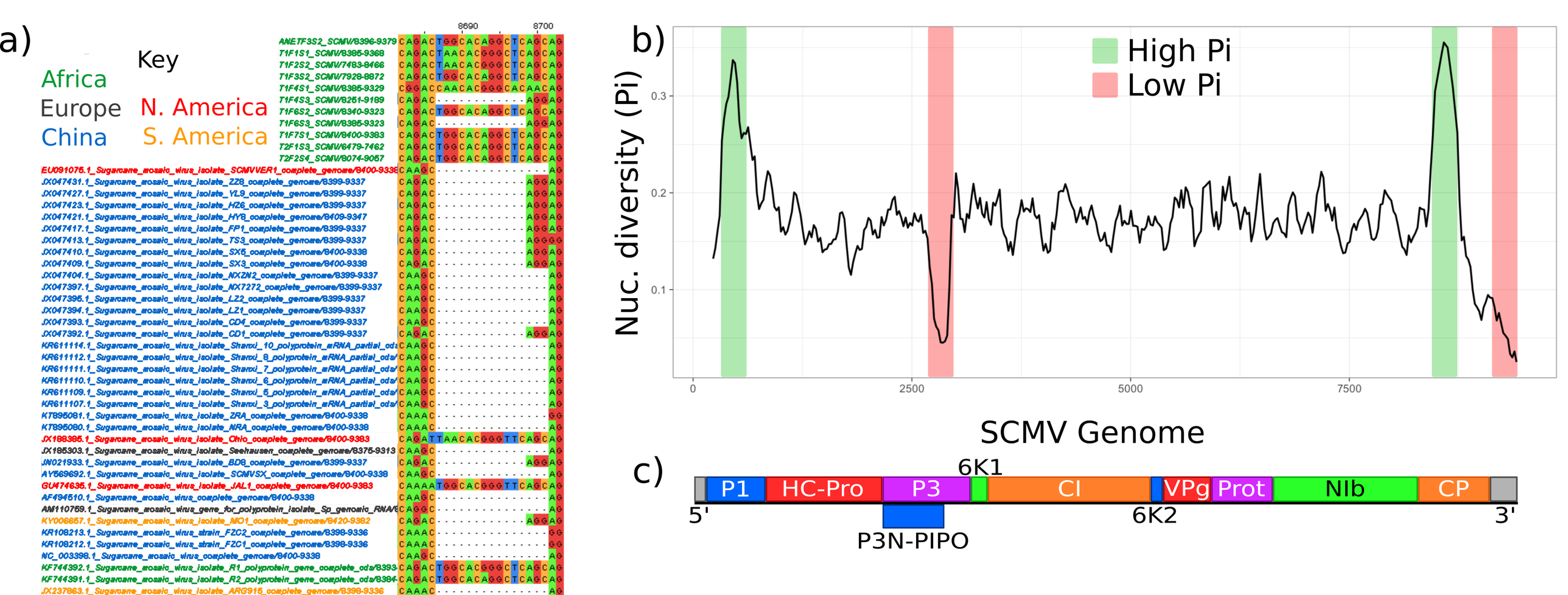
**Figure 1. *Sugarcane mosaic virus* is a worldwide crop pathogen** a) Distribution of *Sugarcane mosaic virus* (SCMV), using country incidence data from CABI. Map generated using data from the maps package with ggplot2 in R v3.4.1[66]. Lat - latitude; long - longitude. b) SCMV genome structure, showing final protein products (initially transcribed as a polyprotein), and previously reported potyviral recombination hot-spots.

**Figure 2. Structural and sequence variation in *Sugarcane mosaic virus*** a) Nucleotide alignment of *Sugarcane mosaic virus* (SCMV) genomes, showing insertion/deletion polymorphism in the coat protein gene. b) Nucleotide (nuc.) diversity across the SCMV genome, with a window size of 100 bp and step size of 25 bp. c) SCMV genome structure, showing final protein products, aligned with diversity graph in b).

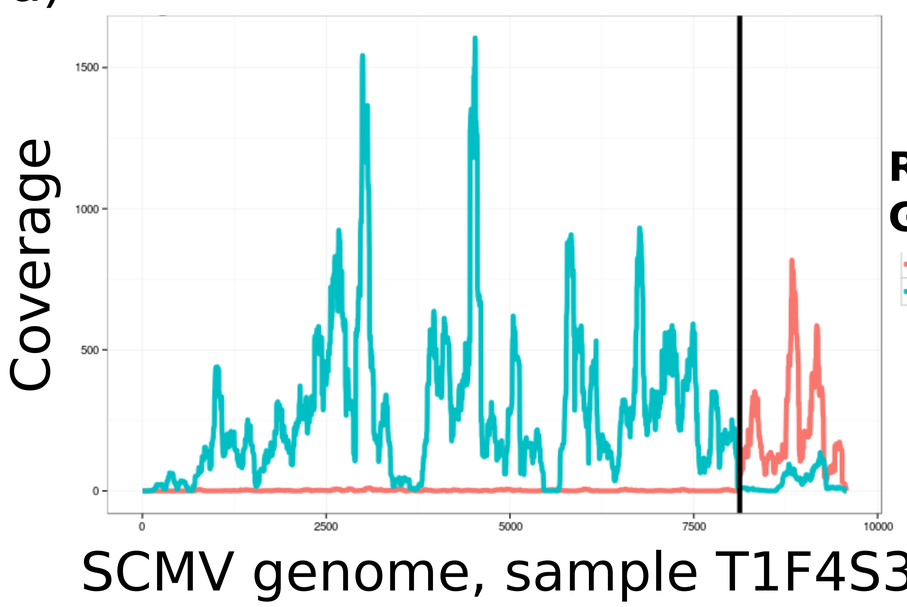
**Figure 3. Widespread recombination between the ancestors of geographically separated *Sugarcane mosaic virus*** a) Bowtie2 alignment of sample T1F4S3 to two divergent *Sugarcane mosaic virus* (SCMV) reference genomes (JN021933.1 and KF744390.1). Reads are assigned to the reference with the best alignment, and the rapid switch in alignment preference is indicated by the black line. b) Splits network of SCMV genomes, distances calculated with uncorrected P, and network generated by neighbour-net in Splitstree V4.6. The reticulate network indicates conflicting phylogenetic signals within the alignment, suggesting recombination. c) Network showing recombination events within and between geographic regions, predicted by Recombination Detection Programme 4 (RDP4). Inter-region recombination refers to recombination events in which the donor (parent) and receiver (child) isolates are from different regions. Intra-region events refer to those in which the parent and child are from the same region. Wider arrows correspond to more recombination events. d) Network showing RDP4-predicted recombination events between individual isolates, with node size determined by the number of recombination events. In b), c), and d) colour indicates geographic region.

**Figure 4. Recombination hot- and cold-spots in the *Sugarcane mosaic virus* genome** a) Recombination events across the *Sugarcane mosaic virus* (SCMV) genome in a 200 bp window. b) P-value distribution for recombination frequency across SCMV genome. Grey ribbons show the local 95 % and 99 % confidence intervals as determined by permutation test. c) SCMV genome structure, showing final protein products and recombination hot- and cold-spots, aligned with graphs in a) and b). Sections used for statistical analysis of conflicting phylogenetic signals are shown. C = control.

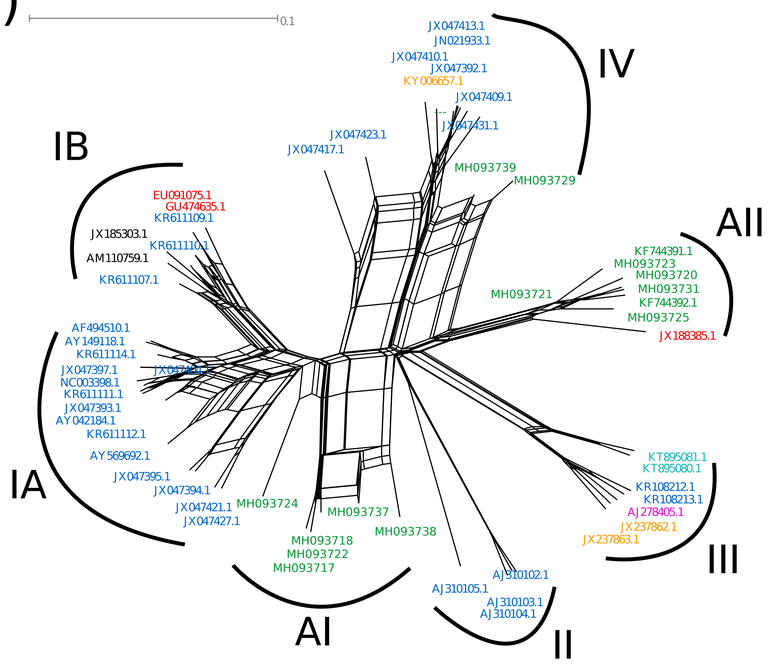

a)
50
0
−50
Lat.
−100
0
100
200
Long.
SCMV presence
b)
P1
HC-Pro
P3
6K1
CI
VPg
Prot
NIb
CP
5'
3'
6K2
P3N-PIPO
Potyvirus recombination hotspots


a)
Key
Africa
Europe
China
N. America
S. America
8690
8700
ANETF3S2_SCMV|8396-9379
T1F1S1_SCMV|8385-9368
T1F2S2_SCMV|7483-8466
T1F3S2_SCMV|7928-8872
T1F4S1_SCMV|8385-9329
T1F4S3_SCMV|8251-9189
T1F6S2_SCMV|8340-9323
T1F6S3_SCMV|8385-9323
T1F7S1_SCMV|8400-9383
T2F1S3_SCMV|6479-7462
T2F2S4_SCMV|8074-9057
EU091075.1_Sugarcane_mosaic_virus_isolate_SCMVVER1_complete_genome/8400-9338
JX047431.1_Sugarcane_mosaic_virus_isolate_ZZ8_complete_genome/8399-9337
JX047427.1_Sugarcane_mosaic_virus_isolate_YL9_complete_genome/8399-9337
JX047423.1_Sugarcane_mosaic_virus_isolate_HZ6_complete_genome/8399-9337
JX047421.1_Sugarcane_mosaic_virus_isolate_HYB_complete_genome/8409-9347
JX047417.1_Sugarcane_mosaic_virus_isolate_FP1_complete_genome/8399-9337
JX047413.1_Sugarcane_mosaic_virus_isolate_TS3_complete_genome/8399-9337
JX047410.1_Sugarcane_mosaic_virus_isolate_SX5_complete_genome/8400-9338
JX047409.1_Sugarcane_mosaic_virus_isolate_SX3_complete_genome/8400-9338
JX047404.1_Sugarcane_mosaic_virus_isolate_NXZN2_complete_genome/8399-9337
JX047397.1_Sugarcane_mosaic_virus_isolate_NX7272_complete_genome/8399-9337
JX047395.1_Sugarcane_mosaic_virus_isolate_LZ2_complete_genome/8399-9337
JX047394.1_Sugarcane_mosaic_virus_isolate_LZ1_complete_genome/8399-9337
JX047393.1_Sugarcane_mosaic_virus_isolate_CD4_complete_genome/8399-9337
JX047392.1_Sugarcane_mosaic_virus_isolate_CD1_complete_genome/8399-9337
KR611114.1_Sugarcane_mosaic_virus_isolate_Shanxi_10_polyprotein_mRNA_partial_cds/
KR611112.1_Sugarcane_mosaic_virus_isolate_Shanxi_8_polyprotein_mRNA_partial_cds/
KR611111.1_Sugarcane_mosaic_virus_isolate_Shanxi_7_polyprotein_mRNA_partial_cds/
KR611110.1_Sugarcane_mosaic_virus_isolate_Shanxi_6_polyprotein_mRNA_partial_cds/
KR611109.1_Sugarcane_mosaic_virus_isolate_Shanxi_5_polyprotein_mRNA_partial_cds/
KR611107.1_Sugarcane_mosaic_virus_isolate_Shanxi_3_polyprotein_mRNA_partial_cds/
KT895081.1_Sugarcane_mosaic_virus_isolate_ZRA_complete_genome/8400-9338
KT895080.1_Sugarcane_mosaic_virus_isolate_NRA_complete_genome/8400-9338
JX188385.1_Sugarcane_mosaic_virus_isolate_Ohio_complete_genome/8400-9383
JX185303.1_Sugarcane_mosaic_virus_isolate_Seehausen_complete_genome/8375-9313
JN021933.1_Sugarcane_mosaic_virus_isolate_BD8_complete_genome/8399-9337
AY569692.1_Sugarcane_mosaic_virus_isolate_SCMVSX_complete_genome/8400-9338
GU474635.1_Sugarcane_mosaic_virus_isolate_JAL1_complete_genome/8400-9383
AF494510.1_Sugarcane_mosaic_virus_complete_genome/8400-9338
AM110759.1_Sugarcane_mosaic_virus_gene_for_polyprotein_isolate_Sp_genomic_RNA/8
KY006657.1_Sugarcane_mosaic_virus_isolate_MO1_complete_genome/8420-9382
KR108213.1_Sugarcane_mosaic_virus_strain_FZC2_complete_genome/8398-9336
KR108212.1_Sugarcane_mosaic_virus_strain_FZC1_complete_genome/8398-9336
NC_003398.1_Sugarcane_mosaic_virus_complete_genome/8400-9338
KF744392.1_Sugarcane_mosaic_virus_isolate_R1_polyprotein_gene_complete_cds/8393-
KF744391.1_Sugarcane_mosaic_virus_isolate_R2_polyprotein_gene_complete_cds/8384-
JX237863.1_Sugarcane_mosaic_virus_isolate_ARG915_complete_genome/8398-9336
b)
High Pi
Low Pi
Nuc. diversity (Pi)
0.3
0.2
0.1
0
2500
5000
7500
SCMV Genome
c)
6K1
P1
HC-Pro
P3
CI
VPg
Prot
NIb
CP
5'
6K2
3'
P3N-PIPO

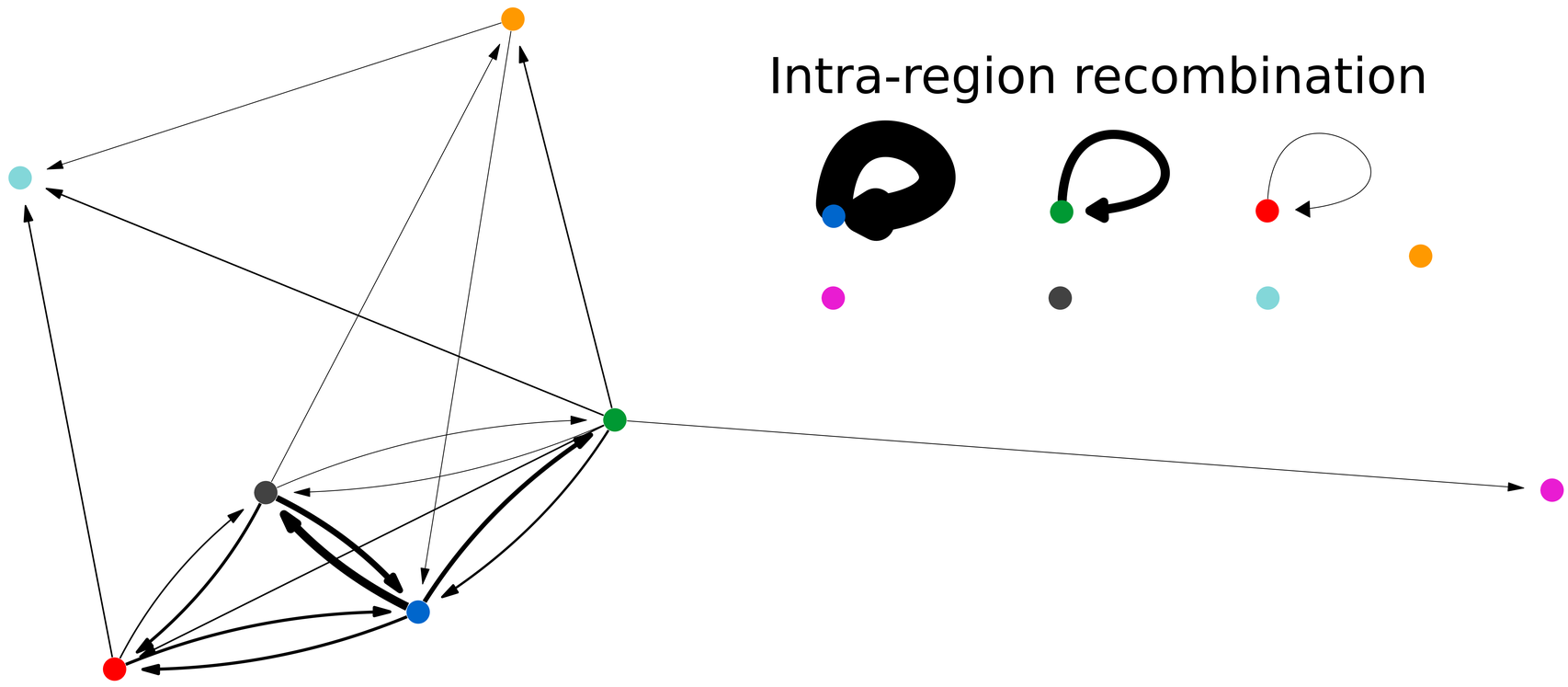
a)




b)




c)



Inter-region recombination

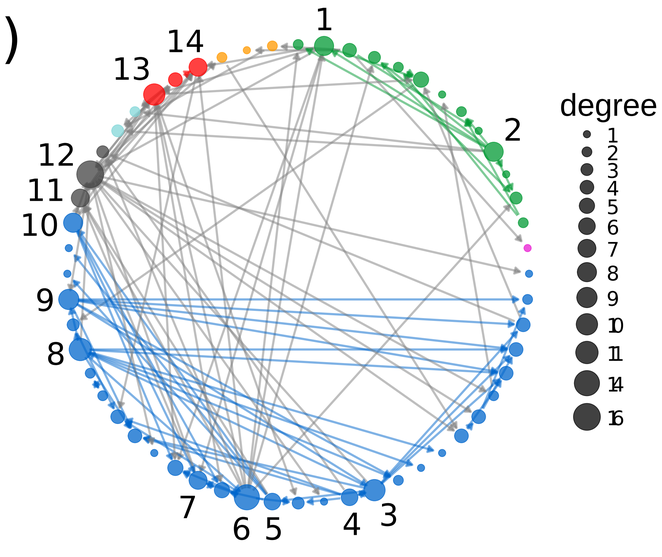
Intra-region recombination



## Key

- Africa
- Australia
- China
- Europe
- Iran
- N. America
- S. America
- Recombinant fragment donor
- Recombinant fragment receiver
- 1 recombination event
- 48 recombination events

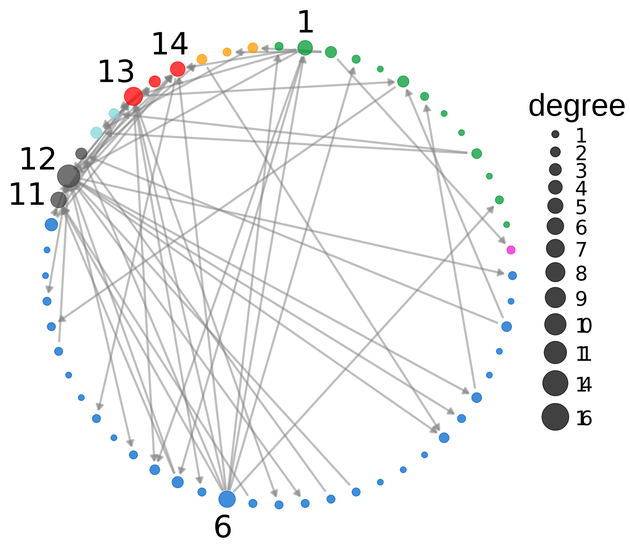
d)



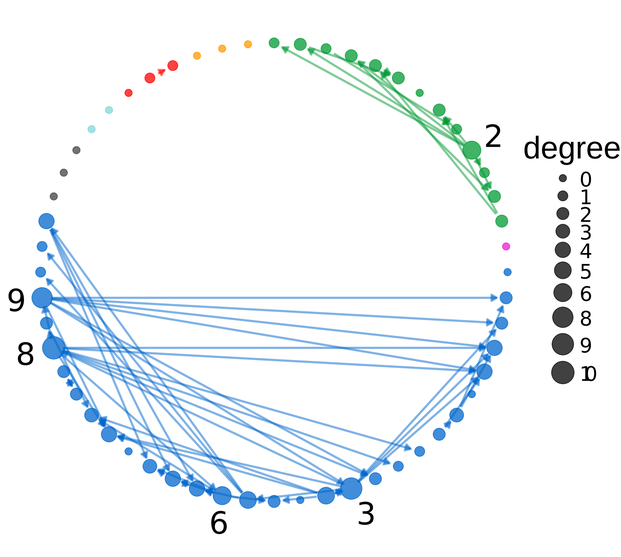

**Connections > 5**

| | |
|---|---|
| 1 MH093718 | 8 NC003398 |
| 2 MH093731 | 9 AJ310105 |
| 3 JX047394 | 10 AJ310102 |
| 4 JX047393 | 11 JX185303 |
| 5 KR611112 | 12 AM110759 |
| 6 KR611111 | 13 EU091075 |
| 7 KR611109 | 14 GU474635 |

Total recombination by isolate




Inter-region recombination




Intra-region recombination

a)

Recomb. events

Global 99%

Global 95%

b)

log(Pval)/-log(Pval)

MH093718 genome position

**Key**

- Local hot-spot
- Local cold-spot
- 95 % range
- 99 % range

c)